BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 10 2006, pages 1282–1283
doi:10.1093/bioinformatics/btl099

Data and text mining

BicAT: a biclustering analysis toolbox

Simon Barkow[1,*], Stefan Bleuler[1], Amela Prelić[1], Philip Zimmermann[2] and Eckart Zitzler[1]

[1]Reverse Engineering Group: Computer Engineering and Networks Laboratory and [2]Institute for Plant Sciences, Swiss Federal Institute of Technology Zurich, ETH Zentrum, 8092 Zurich, Switzerland

Received on July 27, 2005; revised on February 21, 2006; accepted on March 13, 2006
Advance Access publication March 21, 2006
Associate Editor: Thomas Lengauer

ABSTRACT

Summary: Besides classical clustering methods such as hierarchical clustering, in recent years biclustering has become a popular approach to analyze biological data sets, e.g. gene expression data. The Biclustering Analysis Toolbox (BicAT) is a software platform for clustering-based data analysis that integrates various biclustering and clustering techniques in terms of a common graphical user interface. Furthermore, BicAT provides different facilities for data preparation, inspection and postprocessing such as discretization, filtering of biclusters according to specific criteria or gene pair analysis for constructing gene interconnection graphs. The possibility to use different biclustering algorithms inside a single graphical tool allows the user to compare clustering results and choose the algorithm that best fits a specific biological scenario. The toolbox is described in the context of gene expression analysis, but is also applicable to other types of data, e.g. data from proteomics or synthetic lethal experiments.

Availability: The BicAT toolbox is freely available at http://www.tik.ee.ethz.ch/sop/bicat and runs on all operating systems. The Java source code of the program and a developer's guide is provided on the website as well. Therefore, users may modify the program and add further algorithms or extensions.

Contact: barkow@tik.ee.ethz.ch

Microarray technology has become a central tool in biological research, and the identification of gene groups with similar expression patterns represents a key step in the analysis of gene expression data. Traditional clustering algorithms partition an expression matrix into submatrices that extend over the whole set of conditions, giving all conditions equal weight. For specific biological questions, though, the assumption that all genes behave similarly in all conditions, may be too restrictive. To account for this, biclustering approaches carry out the grouping in both dimensions simultaneously: genes and conditions. This allows to find subgroups of genes that show the same response under a subset of conditions, e.g. if a cellular process is only active under these conditions. Furthermore, if a gene participates in multiple pathways that are differentially regulated, one would expect this gene to be included in more than one cluster; this cannot be achieved by traditional clustering.

Several biclustering algorithms have been proposed in the literature, each of which has strengths and weaknesses for the application in different biological scenarios (Madeira and Oliveira, 2004). A recent comparative study has shown that there are significant differences in performance among biclustering approaches, depending on the biological problem that is examined (Prelic *et al*., 2006). Since every algorithm is subject to a specific mathematical problem formulation, it cannot be expected that a single approach is well-suited for all scenarios. Accordingly, it can be useful in practice to try different approaches and to choose that algorithm that delivers the best results. Although implementations are available for some of the proposed biclustering algorithms, each program may be accompanied by a different user interface and use different input and output formats, which in turn makes the application of several methods a time-consuming task. Desirable is a software tool that offers different biclustering approaches within a common framework—to our best knowledge, such a tool has not been available so far. BicAT tries to fill this gap and provides the following functionality:

- Data handling: Tree-structured data handling that allows (1) access to all analysis steps and (2) data export of biclustering and filtering results
- Data preprocessing: Normalization ($\log_2$, mean centric) and discretization
- Clustering: Five biclustering algorithms and two traditional clustering algorithms
- Data visualization: Heatmap and profile visualization of biclusters
- Postprocessing: Analysis of gene pair occurrence to derive gene interconnection graphs

In the following, the main characteristics of the tool will be described—for each of the above aspects separately.

Data handling. All data, be it entire gene expression matrices loaded from external files or sets of submatrices generated by specific biclustering algorithms, are organized in a tree structure that is depicted in the left panel of the graphical user interface, cf. Figure 1. The loaded datasets form the top hierarchy of the tree. The second hierarchy is built up by the clustering, search and filter procedures with the generated lists of biclusters. This structure allows to access every performed analysis step, which is particularly useful, e.g. for the comparison of different clustering runs. Every search and filter operation produces a new list of biclusters that can be further examined in the next analysis steps or exported to a file.

Data preprocessing. The input data file can be any tab-separated text file including annotations of genes and conditions. The loaded gene expression data can then be transformed by means of normalization and discretization. For normalization, the $\log_2$ or the mean centric of the original values can be computed; discretization can be performed with regard to upregulated, downregulated or complementary expression patterns on the basis of a user-defined threshold.

*To whom correspondence should be addressed.

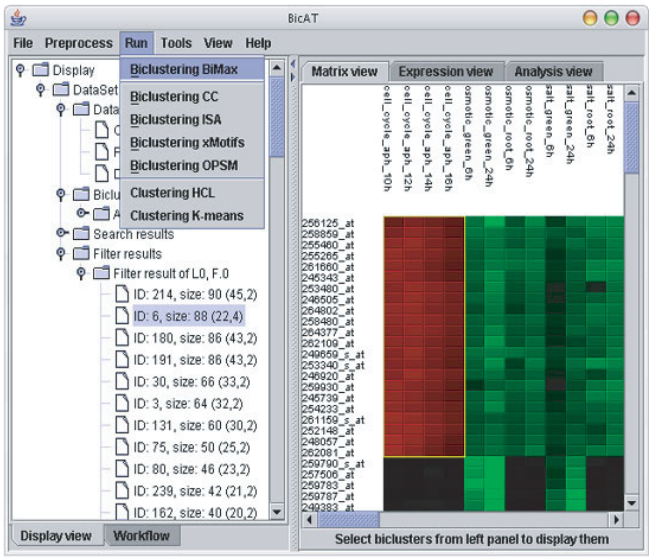

Fig. 1. Graphical user interface of the BicAT software. The window on the left hand side displays the loaded datasets and the conducted analysis steps in a tree structure. The panel on the right shows the heatmap view of an expression matrix with a selected bicluster framed in yellow color.

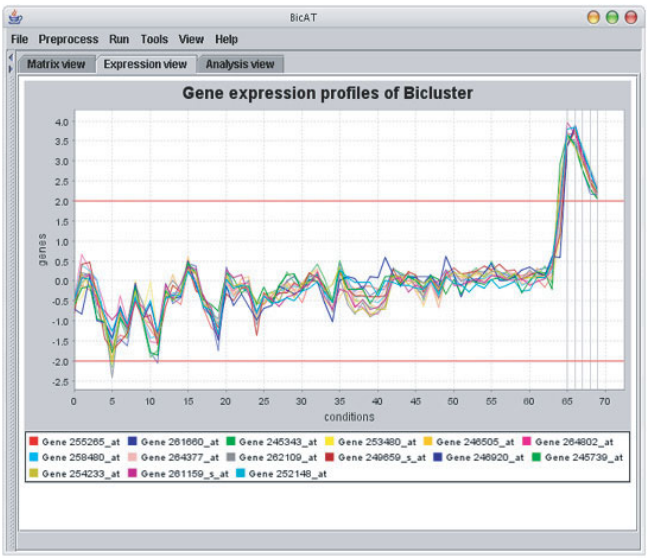

Fig. 2. View of the expression profiles of genes from the selected bicluster. The colored curves stand for single genes in the cluster. The upright black bars on the right mark the conditions that are included in the bicluster. The red lines mark the discretization threshold.

Clustering methods. BicAT implements the following biclustering methods: (1) Cheng and Church's algorithm which is based on a mean squared residue score (Cheng and Church, 2000); (2) the Iterative Signature Algorithm which searches for submatrices representing fix points (Ihmels *et al*., 2002, 2004); (3) the Order-preserving Submatrix Algorithm which tries to identify large submatrices for which the induced linear order of the columns is identical for all rows (Ben-Dor *et al*., 2003); (4) the xMotif algorithm, an iterative search method which seeks biclusters with quasi-constant expression values (Murali and Kasif, 2003); (5) Bimax, an exact biclustering algorithm based on a divide-and-conquer strategy that is capable of finding all maximal bicliques in a corresponding graph-based matrix representation (Prelic *et al*., 2006). In addition, two standard clustering procedures, namely hierarchical clustering and *K*-means clustering, are included.

Visualization. The expression matrix is displayed as a heatmap. Annotations of the conditions run along the top, annotations of the genes are listed on the left hand side. A mouse click on any of the rectangles in the heatmap shows the annotations for a specific data point. Biclusters can be visualized in two different ways: (1) within the heatmap or (2) as a collection of gene expression profiles. As to the first possibility, the heatmap is rearranged in such a way that those genes and conditions that define the bicluster under consideration appear in the upper left corner of the map, Figure 1. Alternatively, the expression view of a bicluster, Figure 2, displays the profiles of those genes that are grouped within a bicluster. Here, for each gene a colored line connects the expression values for the different conditions. Note that the expression view shows all conditions; conditions that are included in the bicluster are marked with upright bars. With the help of the expression profile, the biologist can evaluate the relevance of a specific bicluster.

Postprocessing. For further investigations, BicAT offers the possibility of a gene pair analysis, which summarizes the outcome of a biclustering as a whole. In particular, for each pair of genes it is calculated how often these genes occur together in the same bicluster. This number of co-occurrence indicates which genes may be functionally related. The resulting gene–gene matrix with the according counts can be exported for further visualization and for the derivation of gene interconnection graphs with external tools, e.g. BioLayout by EBI UK (Enright and Ouzounis, 2001).

ACKNOWLEDGEMENTS

The authors would like to thank all members of the Reverse Engineering Group at ETH Zurich for valuable discussions and suggestions. A.P., S.B., and P.Z. have been supported by the SEP program at ETH Zürich under the Poly Project TH-8/02-2. Simon Barkow has been supported by the EU Marie Curie research training network SY-STEM.

Conflict of Interest: none declared.

REFERENCES

Ben-Dor,A. *et al*. (2003) Discovering local structure in gene expression data: the order-preserving submatrix problem. *J. Comput. Biol*., **10**, 373–384.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **8**, 93–103.

Enright,A.J. and Ouzounis,C.A. (2001) BioLayout—an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, **17**, 853–854.

Ihmels,J. *et al*. (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet*., **31**, 370–377.

Ihmels,J. *et al*. (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics*, **20**, 1993–2003.

Madeira,S. and Oliveira,A. (2004) Biclustering algorithms for biological data. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **1**, 24–45.

Murali,T.M. and Kasif,S. (2003) Extracting conserved gene expression motifs from gene expression data. *Pac. Symp. Biocomput*., **8**, 77–88.

Prelic,A. *et al*. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.